Journal of Engineering and Applied Research (JEAR)

National University of Skills

Summer & Autumn 2024, Vol. 1, No. 2, p. 101-113

Journal Homepage: https://jear.tvu.ac.ir/

:10.48301/JEAR.2024.471677.1034

An X-vector-based Speaker Recognition in Persian

Fateme Shahbakhti[1*], Maryam Moradi-Shabestari[2], Zeinab Ghasemi-Naraghi[3]

[1]Department of Electrical and Computer Engineering, Faculty of Shariaty, Skill National University (NUS), Tehran, Iran.
[2]Electrical and Computer Engineering Department, Tehran University, Tehran, Iran.
[3]Computer Engineering Department, Amir Kabir University of Technology, Tehran, Iran.

ARTICLE INFO

Article Type:
Original Research

Received: 08.14.2024
Revised: 08.27.2024
Accepted: 09.03.2024

Keyword:
Deep Neural Networks
Speaker Recognition
X-Vector
Persian Language

***Corresponding Author:**
Fateme Shahbakhti
Email:
fshahbakhti7459@gmail.com

ABSTRACT

In this paper, a text-independent speaker recognition system in Persian was implemented by deep neural networks. The x-vector technique based on Time Delay Neural Network (TDNN) was used to extract the embeddings from speech signals. This method has attracted the attention of researchers due to noise robustness and high performance. Data augmentation and noise addition were used to improve system performance. The PLDA classifier was used to recognize the speaker. Previous research in the field of "speaker recognition in Persian" is limited. In the present study, the network was trained on the Persian part of the CommonVoice dataset. According to the error analysis, non-speech parts of an utterance decrease the accuracy of speaker recognition. Thus, the non-speech parts were removed by a Convolutional Recurrent Deep Neural Network (CRDNN). The accuracy of speaker recognition and verification in CommonVoice was 95.24% and 95.56%, respectively. The Equal Error Rate (EER) evaluation metric of the speaker verification system was 4.72%. The attendance monitoring system was developed as one of the applications of the speaker recognition system. System accuracy for 12 and 15 seconds of collected data (including data from 16 women and 12 men) was 98.92% and 100%, respectively.

E-ISSN: 3060-6357

Introduction

Biometric authentication is a security process based on the unique biological characteristics of a person such as their face, voice, iris, and fingerprints. Speaker recognition consists of two tasks: speaker identification and speaker verification. The purpose of speaker identification is to identify an unknown speaker among the known speakers. This operation is a one-to-many matching problem, meaning that the utterance from the speaker is compared to all available patterns to identify the speaker. Speaker verification is the process of accepting or rejecting an identity. That is, the speaker's voice is compared with the claimed speaker pattern, and if the patterns are similar, the claim is accepted; otherwise, it is rejected [1]. This task is mostly used in security and access control applications. This system has less computational complexity due to fewer comparisons.

Speaker recognition systems are text-dependent or text-independent. In text-dependent mode, the speaker expresses only limited fixed words or sentences that have been used in the training phase. However, in the text-independent mode, the utterances in the test and training signals are different. Text-dependent speaker recognition with short speeches in the test and training phases has good performance. In text-independent recognition, longer speeches are needed for training and testing. In addition, noisy environments have a greater negative impact on the performance of text-independent systems [1].
Before deep learning methods in the field of speaker recognition, the i-vector was widely used as a state-of-the-art method [2]. Today, deep neural network (DNN) methods have been successful in extracting distinguishing features. X-vector is an advanced extraction method based on DNN. This method has better performance than the i-vector, particularly in short speeches [3]. This method is known as a high-performance robust method in noisy environments [4].

A great deal of research has been done on speaker recognition, most of which is in English. Since deep learning methods have gained attention in this area and one of the main challenges in training deep networks is finding an appropriate and large dataset, large open-source datasets such as VoxCeleb [5], LibriSpeech [6], and NIST-SRE 2016 [7] for speaker recognition have been made available in English. The focus of this work is on recognizing the speaker in the Persian language in a practical way in a real environment. The only large-scale open-source database in Persian is CommonVoice. In this paper, a text-independent speaker recognition application is implemented for the Persian part of the CommonVoice dataset.

Speaker recognition can be used in security work, health monitoring [8], and attendance. The purpose of the research was to develop an attendance monitoring application in Persian by combining the best methods of feature

extraction and classification and applying the pre-processing at a suitable speed. Combining the components of removing silence, extracting features and classification, and fine-tuning the network by Persian voices was carried out to adapt it for the intended use. The x-vector embeddings were extracted by the time delay neural network (TDNN). The TDNN model was finetuned with VoxCeleb1+2 initial weights to transfer learning on the CommonVoice dataset. In the last layer of the network, embeddings were classified by PLDA. Error analysis led to a better understanding of the weaknesses of the model in prediction. The main weaknesses of the proposed model were noise and non-speech parts of the utterance. Therefore, noise and silence removal methods were evaluated to improve accuracy. The CRDNN-based silence removal method was successful.
The remainder of the present paper is organized as follows. Section 2 reviews the existing literature in the field of speaker recognition. Section 3 describes the proposed method, which includes preprocessing, x-vector feature extraction, and PLDA classification. The experimental details consisting of dataset details, implementation, and various experiments based on error analysis are explained in Section 4. The results are presented in the final section.

Related works

This section provides a history of speaker recognition methods. The following is an overview of the most important research in this field. In the past, traditional methods were used to recognize the speaker. In 1995, Reynolds et al. used Gaussian Mixture Models (GMM) to develop a text-independent and robust speaker recognition system. They used telephone data from 49 speakers and achieved 80.8% accuracy [9]. Then, Reynolds et al. showed that the GMM-UBM model was effective in speaker recognition tasks in 2000[10]. In 2007, Kenny et al. presented the Joint Factor Analysis (JFA) method based on GMM to solve the intersession variability and channel mismatches problem [11]. In 2011, Dahak et al. proposed a JFA-based i-vector system. I-vector uses the Total Variability Matrix (T) to create a representation of speaker and channel information [2].

With the success of deep learning methods in the fields of machine vision and speaker recognition, scientists' attention was drawn to these methods, particularly for feature extraction. Recently, embedding methods based on deep learning have become prevalent. Snyder et al. proposed a TDNN-based embedding extraction method. This method maps variable-length utterances to fixed-dimension vectors called x-vectors. According to the results, x-vector embedding functions better than i-vector for large-scale training data. The data augmentation technique enhances the x-vector by adding noise and reverberation [12]. Kanagasundaram et al. used deeper x-vector layers with lower dimensions to verify the speaker for a short utterance. For the 5–5 second

data, the EER was 13.35%. This finding is a 14% improvement over the base system [13]. Jahangir et al. proposed a new combination of MFCC and time-based features. The combination of MFCC and time-domain features with DNN was effective in improving accuracy in text-independent speaker systems. Librispeech is a collection of English audiobooks. The EER in the Librispeech dataset for men and women is 0.8% and 0.6%, respectively [14]. Tripathi et al. combined SincNet and X-vector to provide a method to recognize the speaker. The SincNet network extracts frequency-related properties using convolutional layers. In this method, SincNet filters are used in the first layers, and the x-vector is used in the next layers. This approach has led to better network training and increased convergence speed. As a result, the EER for the VoxCeleb1 dataset is 3.56% [15].

Some methods have been proposed for a robust speaker recognition system. Rouvier et al. extended the TDNN by adding a layer. They used the data augmentation technique for robustness by adding music and noise. The EER for speaker verification in the VoxCeleb1E dataset was 1.82% [16]. Wu et al. used a variable auto-encoder (VAE) to learn noisy embedding. The input of the VAE network is an i-vector and an x-vector. They used the manual data augmentation method and GAN-based data augmentation methods in the NIST SRE16 dataset. EER was 4.20% for speaker verification with x-vector input and a PLDA classifier [17]. Mohammadamini et al. used a denoising auto-encoder (DAE) for a robust speaker recognition system. In this method, DAE is used after extracting the x-vector. The EER of the trained model in Voxceleb1+2 for utterances of 2-4 seconds is 6.439% [4]. Taherian et al. used GFCC features to extract the x-vector for speaker verification. They enhance utterance using the gated convolutional recurrent network (GCRN). The EER in NIST SRE 2008 is 8.18% [18]. Kataria et al. used a feature-domain-based denoising method to improve the speaker verification system. They used two x-vector networks called ETDNN and FTDNN to increase domain features. They also used Deep Feature Loss to improve the network. EER was 12.40% in the BabyTrain noisy dataset [19].

Limited research has been conducted in the field of speaker recognition in Persian. DeepMine is a large dataset including Persian and English languages. This dataset is used in speaker and speech recognition. The i-vector/HMM + RLDA method is applied to the text-dependent part of the dataset. The EER for speaker verification of female and male speakers is 2.20% and 1.33%, respectively [20]. In 2020, the DeepMine dataset was used in the speaker verification challenge with short utterances using the x-vector embedding method with the PLDA classifier. The EER was 6.50% in the Persian part of the dataset [21].

Methodology

The speaker recognition system identifies the speaker using the features of their utterance. In this research, a speaker recognition system was developed for text-independent speaker identification and verification in the Persian language. This system can detect the speaker in real-time. The proposed speaker recognition system consists of pre-processing, feature extraction, and classifier phases. The structure of this system is shown in Figure 1. The details of each section will be explained below.

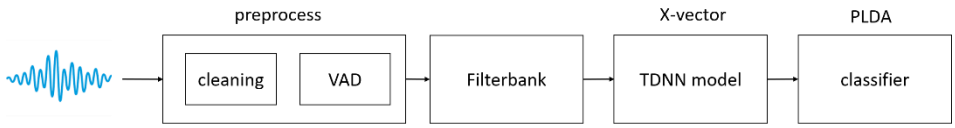

Figure 1. The proposed speaker recognition structure.

Pre-processing

The set of operations applied to raw data before entering the model is called pre-processing. Pre-processing has a considerable effect on model performance. Some methods such as fixed-length and silence removal are used in the pre-processing phase.

Fixed-length: The length of the utterance samples in a dataset varies. Data balance is important for fair training among speakers and better performance. Therefore, the length of all training utterances is a fixed value of 3 seconds.

Silence removal: Every utterance consists of speech and non-speech parts. Non-speech parts are called silence. These parts do not contain useful information from the speaker and can cause the speaker recognition system to malfunction. The voice activity detection (VAD) technique by a neural network is used to identify the parts containing speech in the audio signal. This technique works as a binary classifier. The input is an audio signal. The output is a sequence of 0 and 1. Time frames containing speech are labelled 1, and the ones containing non-speech are labelled 0.

More precisely, the filter bank features are calculated first, and then they are given as input to the CRDNN. This network is a combination of convolutional and recurrent networks. The convolution network extracts high-level local features regardless of location. In the recurrent network, the output of the previous steps is used in the current one, and long-term dependencies are learned. Finally, binary classification is performed by a sigmoid layer. The binary cross-entropy loss function is used to train the network. A brief overview of the VAD system is shown in Figure 2.

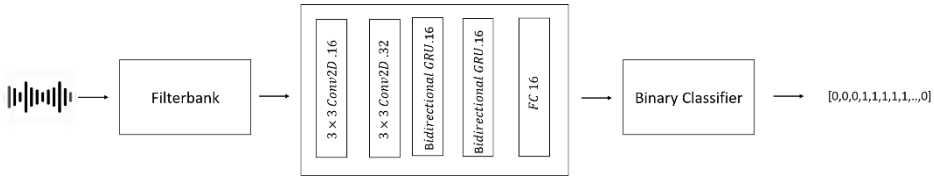

Figure 2. Voice activity detection (VAD) structure.

Data augmentation: Collecting large amounts of labeled data for network training is a costly challenge. One of the solutions used is data augmentation, which improves network performance. In this research, time-masking and frequency-masking, voice speed changes, and adding reverberation and noise to utterances are used.

X-vector

Embedding is extracted by converting high-dimensional utterances into vectors with lower dimensions. This vector maps the features of the speaker to a space with lower dimensions. Among the deep learning methods, the x-vector method has successful performance and high robustness in noisy environments [4]. Thus, in the present study, the x-vector method was used to extract embeddings. The x-vector structure is based on the TDNN model [12] as presented in Table 1. The 24-dimensional filter bank feature with a frame length of 25 ms was calculated for each utterance with a length of 3 seconds. Then, they were provided as input to TDNN. The embedding was extracted at layer segment 6.

Table 1. X-vector structure.

Layer	Layer context	Input*output
Frame 1	[t-2, t+2]	120 * 512
Frame 2	{t-2, t, t+2}	1536 * 512
Frame 3	{t-3, t, t+3}	1536 * 512
Frame 4	{t}	512 * 512
Frame 5	{t}	512 * 1500
State pooling	[0, T)	1500T* 3000
Segment 6	{0}	3000 * 512
Segment 7	{0}	512 * 512
Softmax	{0}	512 * N

Classification

Speaker recognition was performed by classifying embedding. In this research, embeddings with a length of 512 were classified using the probabilistic LDA (PLDA) model. PLDA is a technique that can be considered an improvement over LDA. PLDA is a powerful mechanism for speaker recognition [22]. The PLDA maps data to a subspace with lower dimensions, which increases within-class covariance and decreases between-class covariance.

Test method

In this section, greater details on the implementation of the model are presented. In addition, various experiments and results experienced based on error analysis are reported.

Dataset

CommonVoice [23] is a large microphone dataset. Mozilla has voluntarily collected this dataset in several different languages all over the world. Participants record their voices by reading the sentences displayed on the site. Utterances recorded by other participants are evaluated using a simple voting system. If the recorded utterance first receives two positive votes, the utterance is designated as valid data. If it first receives two negative votes, then it is declared invalid. CommonVoice is the only free dataset in Persian. In this research, a collection of valid Persian utterances without a negative vote was selected. In the process of error analysis, the data with the wrong label was removed. Finally, all speakers with at least 15 utterances were used. The number of these speakers was 1037. To check the generalization and validation of the model and training dataset, a dataset was collected from the voices of Ayeh company employees. This dataset was evaluated for the attendance monitoring system.

Using this dataset to recognize the speaker comes with some challenges such as age and gender imbalance. At the time this research was conducted in July 2021, the Persian data in this dataset included 3655 sentences in 321 hours, of which 282 hours had been confirmed. 39% of Persian-speaking participants were in the age range of 30-39 years, and 33% were in the age range of 19-29 years, of which 75% were men and 7% were women. The second challenge was the different recording conditions for each speaker. These different conditions included a variety of voice recorders and environmental noise. The third challenge was related to the very short length of some utterances.

Implementation

The CommonVoice Persian language dataset was used to train and evaluate the speaker recognition model. Speakers with at least 15 utterance samples were selected, and the selected collection included 1037 speakers. Fifteen selected utterance samples in a ratio of 60, 20, and 20 were used to train, test, and evaluate the x-vector model. The x-vector model with Voxceleb1+2 initial weights was retrained with the CommonVoice dataset using the SpeechBrain toolkit [24]. For the training batch size of 128, the number of epochs was 30, and the learning rate decayed from 0.001 to 0.0001. The NLL-loss function was used for training. As the classification indicated, the PLDA technique with 512 components was used.

To improve performance, data augmentation methods such as time-masking and frequency-masking in the input spectrum and waveform were applied. Noise and reverberation methods were used for robustness and were obtained from the OpenRIR dataset.

Evaluation metrics

Usually, the results of speaker identification tests are compared based on accuracy. These metrics are calculated according to Equation (1).

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}} \quad (1)$$

Equal Rate Error (EER) is another metric that is used to evaluate speaker verification. The amount of equalization between TP and FP is equal to the EER metric.

Experiments

A fundamental step in improving a system's performance is error analysis. Using the confusion matrix, system errors in each class were identified, and then the common features among the classes with the most errors were examined. Finally, appropriate methods to improve the system were applied to these classes.

The base system consisted of an x-vector extractor and a PLDA classifier. The accuracy of this system was 88.98%. Error analysis was used to improve system performance. According to the error analysis, noise and silence in speech cause some bad performances. The impact of removing each case is explained below.

Noise reduction

Noise is an unwanted, annoying, and loud sound. In the current research, the noise of audio signals based on spectral gating was reduced. A spectrogram is an image of the spectral density of signal frequencies at different times. By calculating the spectrogram of a signal and calculating the statistics, the noise threshold was estimated. The mask was calculated by comparing the signal and the noise threshold. Using the filter, the mask was smoothed on frequency and time and applied to the spectrogram. This is a static noise reduction technique. In the non-static method, the estimated noise threshold is constantly updated over time. The results of system evaluation by applying static and non-static noise removal methods to a part of the CommonVoice dataset consisting of 40 speakers are shown in Table 2.

Table 2. Speaker identification results with noise reduction methods 40 speakers in CommonVoice.

Noise Reduction Method	Accuracy (%)
-	98.38
Stationary	97.00
Non-Stationary	97.00

According to the error analysis performed, the CommonVoice dataset contained impulse noise. These noises caused a decrease in system performance. The median filter and Butterworth filter methods were used to remove impulse noise.

The median filter is a noise reduction technique in images and signals. In this technique, the entire input is scrolled with the help of a window, and the middle of each window is given as an output. The median filter has a great effect on removing impulse noise [25].

The Butterworth filter is a signal-processing filter. The input signal is limited to the desired frequency range. A Butterworth filter is commonly used in audio circuits. This filter was invented in 1930 by an engineer named Butterworth [26]. Impulse noise was added to the system's data so that the effect of reducing impulse noise could be observed. Before applying impulse noise reduction methods, all data containing impulse noise was incorrectly classified. After applying the impulse noise reduction methods, it was also classified incorrectly. Therefore, impulse noise reduction methods did not affect improving system performance. The effect of the above noise reduction methods on the audio signal is shown in Figure 3.

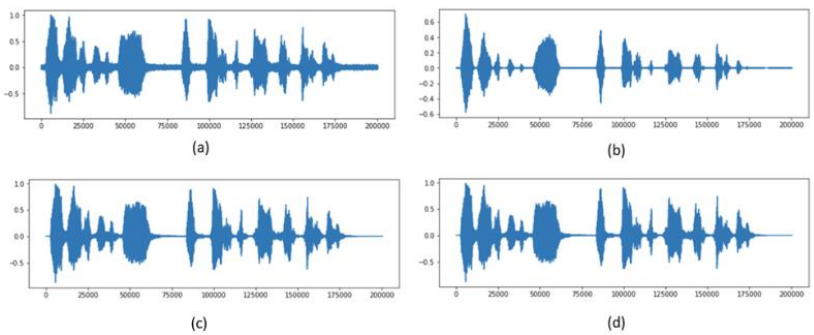

Figure 3. The effect of noise reduction methods on the audio signal a) audio signal with noise; b) audio signal with noise after applying a spectral gate-based method; c) audio signal with noise after applying a middle filter; d) audio signal with noise after applying.

Voice activity detection

Voice activity detection is one method of improving the performance of speaker recognition systems. According to the error analysis, long non-speech parts were detected in the false predictions. By utilizing the voice activity detection method and removing non-speech parts, the accuracy improved by 5.11%. The evaluation results of this experiment with and without applying the VAD preprocess are presented in Table 3.

Table 3. Speaker identification result with VAD on CommonVoice.

Preprocess method	Accuracy (%)
-	88.98
VAD	94.09

Fine-tune

Using trained neural network weights as the initial weights of a new model in the same task domain is a solution to increase the speed of training and solve the problem of a small training dataset. The basic system uses VoxCeleb1+2 weights. In the present study, the TDNN network with the initial weights of VoxCelebe1+2 was fine-tuned on the CommonVoice Persian language dataset. The accuracy of speaker identification with the new network weights is provided in Table 4.

The attendance monitoring system

The attendance monitoring system is one of the applications of the speaker recognition system. A text-independent system was developed for checking the attendance monitoring product of Ayeh individuals. To collect audio data, sentences with different topics were first crawled from websites. Sentences with lengths of less than 10 and more than 20 words were removed. Each speaker recorded 5 random sentences for training with their mobile phone. No repetitive sentences were used in this process. This collection included 16 women and 12 men. System testing was performed with arbitrary sentences. The accuracy of the system with 12 seconds of training data was 98.92%. More data is required to achieve higher accuracy. For example, the proposed system's accuracy with 15 seconds of training data was 100%.

Results and discussion

The accuracy of the identification system for 1037 speakers was 95.24%. The accuracy of the verification was 95.56%, and the EER evaluation metric was 4.72%.

Conclusion

The main challenges of the current study were the presence of noise, silence intervals and its lack of optimization for speaker recognition. Therefore, the designed pre-processing was considered a means of providing a suitable dataset for this research and future research. Furthermore, the lack of dependence of the proposed algorithm on the used data was the key point in the network design so that it can be used under real conditions in the opinion of the research team (using the dataset for speech recognition and speaker recognition in the designed application system).

In the pre-processing step, the CRDNN-based silence removal method was used. In the feature extraction step, the x-vector model with voxceleb1+2 weights was fine-tuned on the CommonVoice dataset. Finally, classification was performed using the PLDA method. The system obtained achieved satisfactory results. Each speaker was identified in real-time. The accuracy of the identification system for 1037 speakers was 95.24%. The performance of speaker verification systems was independent of number. The accuracy of the verification was 95.56%, and the EER evaluation metric was 4.72%. The proposed system was created in the form of an operational product for attendance monitoring. System accuracy for 12 and 15 seconds of collected data (data from 16 women and 12 men) was 98.92% and 100%, respectively.

Disclosure statement and funding

The authors declare no potential conflicts of interest. The present study received no financial support from any organization or institution.

Acknowledgment

The Ayeh Engineering Group supported this research. We gratefully acknowledge the Ayeh Group and thank Mozilla for their open-access dataset.

References

[1] Furui, S. (1996). An Overview of Speaker Recognition Technology. In C-H. Lee, F. K. Soong, & K. K. Paliwal (Eds.), *Automatic Speech and Speaker Recognition: Advanced Topics* (pp. 31-56). Springer. https://doi.org/10.1007/978-1-4613-1367-0_2

[2] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *Institute of Electrical and Electronics Engineers Transactions on Audio, Speech, and Language Processing*, *19*(4), 788-798. https://doi.org/10.1109/TASL.2010.2064307

[3] Okabe, K., Koshinaka, T., & Shinoda, K. (2018, September 2-6). *Attentive statistics pooling for deep speaker embedding* [Conference session]. 2018 International Speech Communication Association, Hyderabad, India. https://doi.org/10.21437/interspeech.2018-993

[4] Mohammad Amini, M., & Matrouf, D. (2021, January 18-21). *Data augmentation versus noise compensation for x-vector speaker recognition systems in noisy environments* [Conference session]. 28th European Signal Processing Conference, Amsterdam, Netherlands. https://doi.org/10.23919/Eusipco47968.2020.9287690

[5] VoxCeleb. (n.d.). *VoxCeleb: Large-scale audio-visual datasets of human speech.* https://mm.kaist.ac.kr/datasets/voxceleb/#downloads

[6] Openslr. (n.d.). *LibriSpeech ASR corpus.* https://www.openslr.org/12

[7] Nist. (2016, August 4). *Speaker Recognition Evaluation 2016*. https://www.nist.gov/system/files/documents/2016/10/07/sre16_eval_plan_v1.3.pdf

[8] Hom, K. L., Beigi, H., & Betti, R. (2022). Application of Speaker Recognition x-Vectors to Structural Health Monitoring. In Z. Mao (Ed.), *Model Validation and Uncertainty Quantification, Volume 3* (pp. 139-148). Springer International Publishing. https://doi.org/10.1007/978-3-030-77348-9_18

[9] Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *Institute of Electrical and Electronics Engineers Transactions on Speech and Audio Processing*, *3*(1), 72-83. https://doi.org/10.1109/89.365379

[10] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, *10*(1-3), 19-41. https://doi.org/10.1006/dspr.1999.0361

[11] Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *Institute of Electrical and Electronics Engineers Transactions on Audio, Speech, and Language Processing*, *15*(4), 1435-1447. https://doi.org/10.1109/TASL.2006.881693

[12] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April 15-20). *X-Vectors: Robust DNN Embeddings for Speaker Recognition* [Conference session]. 2018 Institute of Electrical and Electronics Engineers International Conference on Acoustics, Speech and Signal Processing, Calgary, Alberta, Canada. https://doi.org/10.1109/ICASSP.2018.8461375

[13] Kanagasundaram, A., Sridharan, S., Ganapathy, S., Singh, P., & Fookes, C. (2019, September 15-19). *A study of x-vector based speaker recognition on short utterances* [Conference session]. Proceedings of the 20th Annual Conference of the International Speech Communication Association, Graz, Austria. https://doi.org/10.21437/Interspeech.2019-1891

[14] Jahangir, R., Teh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., Akhtar, M. Z., & Ali, I. (2020). Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network. *Institute of Electrical and Electronics Engineers Access*, *8*, 32187-32202. https://doi.org/10.1109/ACCESS.2020.2973541

[15] Tripathi, M., Singh, D., & Susan, S. (2020). Speaker Recognition Using SincNet and X-Vector Fusion. In L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, & J. M. Zurada (Eds.), *Artificial Intelligence and Soft Computing* (pp. 252-260). Springer International Publishing. https://doi.org/10.1007/978-3-030-61401-0_24

[16] Rouvier, M., Dufour, R., & Bousquet, P. M. (2021, January 18-21). *Review of different robust x-vector extractors for speaker verification* [Conference session]. 28th European Signal Processing Conference, Amsterdam, Netherlands. https://doi.org/10.23919/Eusipco47968.2020.9287426

[17] Wu, Z., Wang, S., Qian, Y., & Yu, K. (2019, September 15-19). *Data Augmentation Using Variational Autoencoder for Embedding Based Speaker Verification* [Conference session]. Proceedings of the 20th Annual Conference of the International Speech Communication Association, Graz, Austria. http://dx.doi.org/10.21437/Interspeech.2019-2248

[18] Taherian, H., Wang, Z. Q., Chang, J., & Wang, D. (2020). Robust Speaker Recognition Based on Single-Channel and Multi-Channel Speech Enhancement. *Institute of Electrical and Electronics Engineers/Association for Computing Machinery Transactions on Audio, Speech, and Language Processing*, *28*, 1293-1302. https://doi.org/10.1109/TASLP.2020.2986896

[19] Kataria, S., Nidadavolu, P. S., Villalba, J., Chen, N., García-Perera, P., & Dehak, N. (2020, May 4-8). *Feature Enhancement with Deep Feature Losses for Speaker Verification* [Conference session]. 2020 Institute of Electrical and Electronics Engineers International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain. https://doi.org/10.1109/ICASSP40776.2020.9053110

[20] Zeinali, H., Sameti, H., & Stafylakis, T. (2018, June 26-29). *DeepMine Speech Processing Database: Text-Dependent and Independent Speaker Verification and Speech Recognition in Persian and English* [Conference session]. The Speaker and Language Recognition Workshop, Les Sables d'Olonne, France. http://dx.doi.org/10.21437/Odyssey.2018-54

[21] Khoa, T. D., & Tsai, T. H. (2020, October 30-31). *A Text-Independent Speaker Verification for SdSV Challenge 2020* [Conference session]. 2020 Institute of Electrical and Electronics Engineers 5th International Conference on Computing Communication and Automation, Greater Noida, India. https://doi.org/10.1109/ICCCA49541.2020.9250773

[22] Khosravani, A., & Homayounpour, M. M. (2017). A PLDA approach for language and text independent speaker recognition. *Computer Speech & Language*, *45*, 457-474. https://doi.org/10.1016/j.csl.2017.04.003

[23] CommonVoice. (2021). *Datasets*. https://commonvoice.mozilla.org/en/datasets

[24] Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J-C., Yeh, S-L., Fu, S-W., Liao, C-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., De Mori, R., & Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. *arXiv*, 1-34. https://doi.org/10.48550/arXiv.2106.04624

[25] Sagayam, K. M., Bruntha, P. M., Sridevi, M., Renith Sam, M., Kose, U., & Deperlioglu, O. (2021). A cognitive perception on content-based image retrieval using an advanced soft computing paradigm. In T. Gandhi, S. Bhattacharyya, S. De, D. Konar, & S. Dey (Eds.), *Advanced Machine Vision Paradigms for Medical Image Analysis* (pp. 189-211). Academic Press. https://doi.org/10.1016/B978-0-12-819295-5.00007-X

[26] Butterworth, S. (1930). On the theory of filter amplifiers. *Wireless Engineer*, *7*(6), 536-541. https://www.changpuak.ch/electronics/downloads/On_the_Theory_of_Filter_Amplifiers.pdf